区域医疗健康平台中检验检查指标的标准化算法研究

张佳影[1] 王祺[1] 张知行[1] 阮彤[1] 张欢欢[1] 何萍[2]

[1]（华东理工大学 上海市 200237）

[2]（上海申康医院发展中心 上海市 200041）

（zhangjy_ecust@163.com）

Lab Indicator Standardization in a Regional Medical Health Platform

Zhang Jiaying[1], Wang Qi[1], Zhang Zhixing[1], Ruan Tong[1], Zhang Huanhuan[1] and He Ping[2]

[1]（*East China University of Science and Technology, Shanghai 200237*）

[2]（*Shanghai Hospital Development Center, Shanghai 200041*）

Abstract Due to the lack of a complete synonym list for indicator mapping, different hospitals may use different names for the same lab indicator. Lab indicator name discrepancy has greatly affected the medical information sharing and exchange among hospitals. It is becoming increasingly important to standardize the lab indicators. Such a problem can be seen as an entity alignment task to map different indicators into standard ones. However, a lab indicator only involves its name and value, not including any extra properties or contexts which is needed by existing knowledge base (KB) alignment or entity linking methods. More importantly, there exists no available standard KBs to provide standard indicator terms. Therefore, we cannot implement these existing methods directly. To solve the problem, in this paper, we present the first effort to work on lab indicator standardization. We propose a novel standardization method, which firstly cluster the indicators based on their names and abbreviations, and then iteratively employ a binary classification algorithm based on similarity features and partition score features for indicator mapping. Experimental results on the real-world medical data show that the final classification achieves a F1-score of 85.27%, which indicates that our method improves the quality and outperforms state-of-the-art approaches.

Key words regional medical health platform; lab indicator; standardization; clustering; classification

摘要 由于没有完整可用的指标同义词库以进行指标映射，各家医院关于同一检验检查指标的不同称谓，已严重影响到了区域间医疗信息的互联共享，因而需要对检验检查指标进行标准化处理．这可以看作是一个实体对齐问题，但指标只有相应的取值和取值范围，难以像知识库实例匹配那般使用到属性信息，也不似实体链接那般拥有上下文信息，而且不存在一个标准知识库来提供所有指标的标准名称．该文针对以上问题，提出指标标准化算法，先根据指标字面特征进行聚类，再使用相似度特征和分块打分特征迭代地进行二分类映射．实验表明，最终的二分类映射，其 F1-score 可以达到 85.27%，证明了该方法的有效性．

关键词 区域医疗健康平台；检验检查指标；标准化；聚类；分类

中图法分类号 TP391

基金项目：国家自然科学基金项目（61772201）；国家重点研发计划资助“精准医学研究”重大专项项目（2018YFC0910500）；国家重大新药创制项目（2018ZX09201008）

This work is supported by the National Natural Science Foundation of China (61772201), the National Key R&D Program of China for “Precision medical research ” (2018YFC0910500), and the National Major Scientific and Technological Special Project for “Significant New Drugs Development”（2018ZX09201008）.

通信作者：张欢欢（hzhang@ecust.edu.cn）

随着医疗信息化的不断深入，国内外在现有医疗体系之上相继建立起了区域医疗健康平台。以上海市为例，随着上海市医联工程项目在 2008 年 3 月正式投入使用，上海市建成了包括市内 38 家三级医院的临床诊疗信息共享平台，实现了对患者的基本信息、基本病历资料、住院病案资料、医嘱资料、医疗费用资料、实验室检验检查报告、医学影像检查报告的交换共享，并通过网站等其它辅助系统加强各医院间的协同诊疗。然而，由于历史原因，各家医院关于同一检验检查指标的称谓不尽相同。仅以“血清钠”为例，便有“钠离子浓度”、“NA+”、“动脉血钠”、“血钠(Na)”等 10 多种不同说法。由于目前并没有完整可用的指标同义词库以进行指标映射，这一问题已严重影响到了区域间医疗信息的互联共享。由此，对区域医疗健康平台中检验检查指标做标准化处理，将各家医院的同一指标的不同称谓映射成统一的标准名称，便显得至关重要。然而，由于检验检查指标涉及到大量的医学知识，加之各家医院的指标体系纷繁庞杂，由医学专业人员对其进行人工标准化，需要耗费大量的时间与精力。因此，如何设计一个检验检查指标的标准化算法，便成了关键所在。

检验检查指标的标准化问题，可以看作是一个实体对齐问题，即将医疗健康平台中的候选指标映射到标准指标上。关于实体对齐，目前主要有两类任务，分别是不同知识库中实体间的实例匹配[1][2]，以及文本中实体和知识库实体之间的实体链接[3][4]。前者常利用知识库中实体的属性信息进行实例匹配，后者常利用文本中实体的上下文信息与知识库中实体的属性信息进行实体链接。然而，本文的任务与以上两种任务都不同：检验检查指标存在于电子病历之中，只有相应的取值及取值范围，而不存在属性信息；同时，它也不似文本中实体那般拥有上下文信息；更重要的是，本文任务中并不存在一个标准知识库来提供所有指标的标准名称。也就是说，目前的方法都难以直接适用于本任务。

有鉴于此，针对区域医疗健康平台中的检验检查指标标准化问题，本文提出了一种指标标准化算法框架，首先对指标数据进行预处理，接着利用指标的字面特征，通过基于密度的聚类算法，将不同的指标聚为一个个簇，以缩小指标的对齐范围。然后，为每一个簇确定一个标准名称，并利用二分类算法找出簇内标准名称的同义指标。对于剩下非同义指标，从中筛选出一个新的标准名称，继续利用二分类算法进行同义指标的查找 [1]，如此迭代进行，直到所有簇内均为同义指标或簇内只剩 1 个指标为止。最后，再由医学专业人员对指标对齐结果进行修正处理。实验结果表明，在上海市 8 家三级医院的实验数据集上，最终的二分类映射算法 F1-score 可以达到 85.27%。

1 相关工作

区域医疗健康平台中的检验检查指标标准化问题，可以看作是一个实体对齐问题，即将各家医院的不同指标称谓映射到统一的标准指标上。目前的实体对齐任务基本可以分为两类，分别是不同知识库中实体间的实例匹配，以及文本中实体和知识库中实体之间的实体链接。

许多研究聚焦于知识库实体间的实例匹配，这些研究利用知识库中实体的属性信息进行匹配，它们基本可以分为两类，分别是成对实体匹配方法和集体实体匹配方法。成对实体匹配方法主要有基于传统概率模型的方法、有监督学习的方法、聚类方法和主动学习方法。传统概率模型方法根据属性相似性进行成对实体比较[5][6]，有监督学习方法常使用决策树[1][7][8]、支持向量机[2][9]、集成学习[10][11]等方法进行二分类，聚类方法利用属性相似性进行实体聚类[12][13][14]，主动学习方法通过人机交互不断迭代来训练分类模型[15][16][17]。集体实体匹配方法则将实体的关联实体也纳入考虑，常见的方法有 LDA 方法[18][19]、CRF 模型[13][20]、Markov 逻辑网[21][22]等。

就文本中实体与知识库实体间的实体链接而言，主要有基于概率生成模型的方法[3][23]、基于主题模型的方法[4][24]、基于图的方法[25][26][27][28]和基于深度神经网络的方法[29][30][31][32]。

需要注意的是，本文的研究内容和以上两种研究都不相同：检验检查指标存在于电子病历之中，只有相应的取值和取值范围，难以像知识库实例匹配那般使用到属性信息；同时，它也不似文本中实体那般拥有上下文信息，因而难以使用实体链接的方法；更重要的是，本文任务中并不存在一个标准知识库以提供所有指标的标准名称。

2 指标标准化算法

指标标准化算法的整体流程如图 1 所示。首先，对指标数据进行预处理，实现大小写统一、单位统一和指标参考值提取。接着，利用指标的字面特征，通过基于密度的聚类算法，将不同的指标聚为一个个指

[1] 当然也可以对所有的非同义指标重新进行聚类，如此迭代进行。只是在实际应用时考虑到 38 家医院的不同指标太多，聚类的时间成本很高，本文作为区域医疗健康平台中检验检查指标的标准化算法的初步尝试，暂且迭代使用二分类算法进行标准化

标簇，以缩小指标的对齐范围。然后，为每一个簇确定一个标准名称，并利用二分类算法找出簇内标准名称的同义指标，进行指标映射。对于剩下非同义指标，从中筛选出一个新的标准名称，继续利用二分类算法进行同义指标的查找，如此迭代进行，直到所有簇内均为同义指标或簇内只剩 1 个指标为止。最后，再由医学专业人员对指标对齐结果进行修正处理。

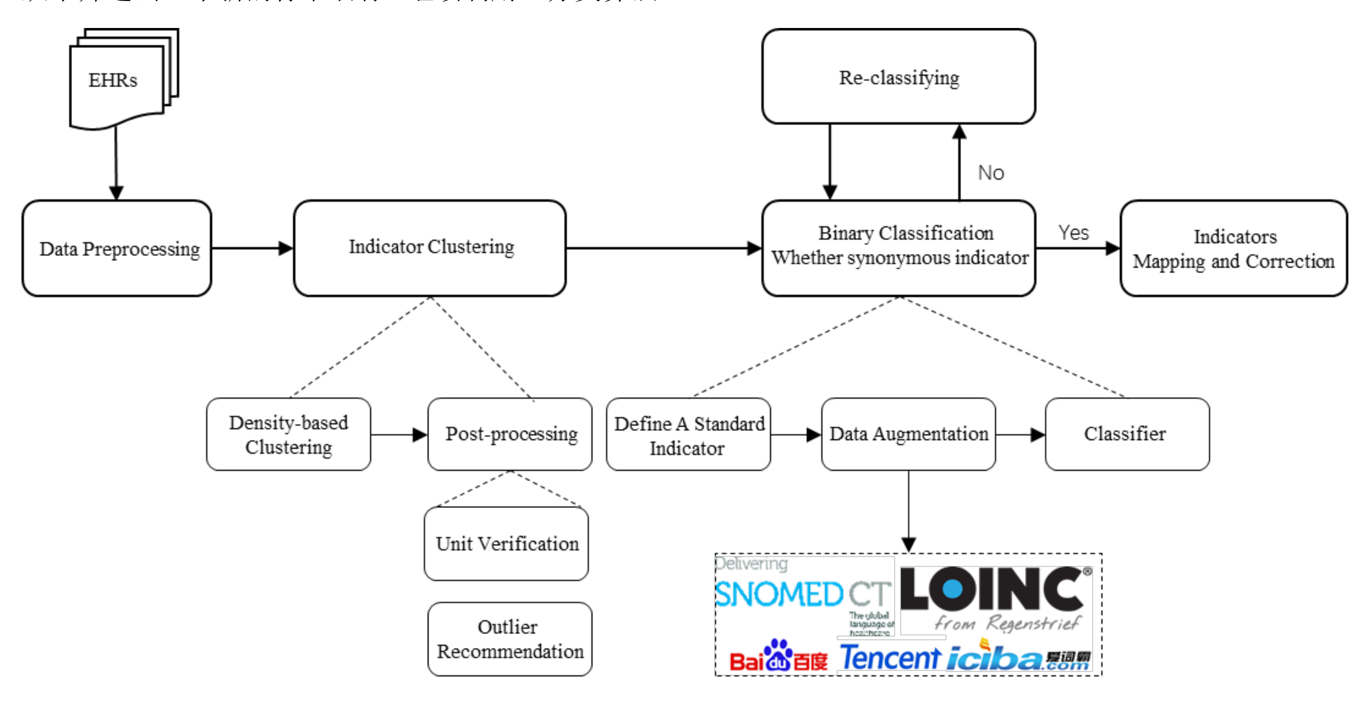


Fig. 1 Overall process of indicator standardization algorithm.

图 1 指标标准化算法整体流程

2. 1 数据预处理

病历中的指标数据，排除选填项 [2]，必填项中主要包括指标名称、缩写、参考值、单位、所属检查项、检查指标结果、异常指标提示等字段。其中，所属检查项因各家医院标准不一、检查指标结果因其取值因病人而异、异常指标提示因不具有指标区分度而失去作为指标标准化特征的意义。因此，可用的字段基本仅限于指标名称、缩写、参考值和单位这 4 项。对指标进行数据预处理，主要是统一指标大小写、统一指标单位，以及提取指标参考值。

2. 2 指标聚类

为缩小指标的对齐范围，本文使用基于密度的聚类算法，将不同的指标聚到一个个指标簇中。基于密度的聚类算法依据样本分布的紧密程度来划分簇，它主要考察样本的可连接性，并在可连接样本的基础上通过不断扩展聚类簇来获得最终结果。

本文基于 DBSCAN[33]算法，使用指标名称及其缩写进行指标聚类。具体来说，给定指标集合 $D=\{x_1, x_2, \ldots, x_n\}$，$x_i=(x_i^{na}, x_i^{ab})$，$i=1, 2, \ldots, n$，其中 x_i^{na} 表示第 i 个指标的指标名称，x_i^{ab} 表示第 i 个指标的名称缩写，定义 ε-邻域及核心对象为：

定义 1 （ε−邻域） 对于 $x_i \in D$,它的 ε-邻域为数据集 D 中与 x_i 的距离不大于 ε 的所有样本，即 $N_\varepsilon(x_i)=\{x_i \in D \mid \mathrm{dist}(x_i, x_j) \leqslant \varepsilon\}$。

定义 2 （核心对象） 如果 x_i 的 ε-邻域内至少包含 $minPts$ 个样本，即 $|N_\varepsilon(x_i)| \geqslant minPts$，那么 x_i 是一个核心对象。

特别地，在确定 ε-邻域时，本文定义了联合距离 $dist_{joint}(x_i,x_j)$：将指标数据 x_i、x_j 分为两部分计算，首先计算 multi-hot 形式（0-1 向量中不同的维度表示不同的汉字）的指标名称 x_i^{na} 、 x_j^{na} 的余弦距离：

$$dist_{cos}(x_i^{na}, x_j^{na}) = 1 - \frac{\sum_{k=1}^{d} x_{i,k}^{na} x_{j,k}^{na}}{\sqrt{\sum_{k=1}^{d} x_{i,k}^{na\,2}} \sqrt{\sum_{k=1}^{d} x_{j,k}^{na\,2}}} \tag{1}$$

然后计算指标缩写 x_i^{ab} 、 x_j^{ab} 的编辑距离：

[2] 实际操作中选填项一般无数据填入。如“LOINC 编码”字段是帮助识别是否是同一个指标的重要特征，然而由于是选填项，实际上没有任何数据填入。

$$dist_{ed}(x_i^{ab},x_j^{ab})=\frac{med(x_i^{ab},x_j^{ab})}{\max(|x_i^{ab}|,|x_j^{ab}|)} \quad (2)$$

其中$|x_i^{ab}|$是指标缩写x_i^{ab}的字符串长度，$med(x_i^{ab},x_j^{ab})$表示由x_i^{ab}经插入、替换、删除操作转成x_j^{ab}所需的最少操作次数。最后，利用调和平均综合两个距离得到联合距离：

$$dist_{joint}(x_i,x_j)=\frac{2\times dist_{cos}(x_i^{na},x_j^{na})\times dist_{ed}(x_i^{ab},x_j^{ab})}{dist_{cos}(x_i^{na},x_j^{na})+dist_{ed}(x_i^{ab},x_j^{ab})} \quad (3)$$

聚类算法从核心对象出发，不断向外扩展，进而生成聚类簇，其伪代码如算法 1 所示。

Algorithm 1: Density-based clustering algorithm

```
Input:(1) Indicators set D={x1, x2, …, xn}
      (2) Neighborhood parameters(ε, minPts)
Output: Cluster partition C={C1, C2, …, Cm}
Initialize the Core Object collection: T=∅
for xi ∈D do
    Determine the Eps-neighborhood: Nε(xi)
    if |Nε(xi)| ≥minPts then
        Add xi to the Core Object set: T =T∪ {xi}
    end
end
Initialize number of clusters: k=0, cluster set: C=∅
unvisited set: P=D
while T≠∅ do
    Record currently not visited collection: P'=P
    Select a core Object o randomly from T;
    Initialize the queue Q= [o]
    Remove o from p,T:P=P-o; T=T-o
    while Q=∅ do
        Take the first sample q in queue Q
        if |Nε(q)| ≥minPts then
            S=P∩Nε(q)
            Q=Q+S
            P=P-S
        end
    end
    k=k+1
    Generate cluster: Ck=P'-P
    T=T−Ck
end
return C
```

需要注意的是，由于聚类是一个无监督的学习过程，它可能存在两个问题：1）聚为一簇的指标实际上医学含义不同，却因为名称相近或缩写相似而被归为一簇；2）有些离群值既不是核心对象，又不能通过核心对象访问，因而没有被聚类。因此，需要对聚类结果进行后处理。

①单位验证。假设同义指标的单位是相同的，那么可以对每一簇指标进行单位验证，将不同单位的指标分离为不同的簇。

②离群值推荐。对于未被聚类的离群值，有两种处理方案，第一种是按距离远近，将离群值分到单位相符且距离最近的那一簇中；第二种是考虑到离群值与其它簇都距离较远，很可能它本身就是一个全新的指标。本文采用第二种处理方案。

2. 3 簇内二分类

即使经过后处理，无监督聚类算法也无法保证簇内的指标皆为同义指标。因此，本文为每一个簇确定一个标准名称，并利用二分类算法将簇内指标划分为标准名称的同义指标和非同义指标两类。特别地，为方便医学专业人员对指标对齐结果进行后处理修正，考虑到标准指标应为最常用的指标，本文以簇内出现频次最多的指标为标准指标。

1）数据增强

由于医学专业人员很难凭空枚举出所有的同义指标，加上有些指标可能会有与名称毫无联系的同义词（如“B 型钠尿肽”和“脑钠素”），因此在数据集生成方面，除由医学专业人员手动标注部分同义指标用于分类器训练之外，本文还利用 SNOMED CT 知识库[34]、LOINC 知识库[35]、百度百科 [3]等途径来抽取标准指标的同义词用于训练。其中，SNOMED CT 知识库为全英文库，目前并无中文版本，因此需要借助百度翻译 [4]、腾讯翻译 [5]、爱词霸翻译 [6]等翻译工具将英文指标翻译为中文指标。需要注意的是，即使对同一个指标，翻译工具也有可能会得到不同的翻译结果，因此翻译本身也是获取同义词的途径之一。表 1 给出了“B 型钠尿肽”经数据增强后的同义指标示例

[3] https://baike.baidu.com/
[4] http://fanyi.baidu.com/
[5] http://fanyi.qq.com/
[6] http://www.iciba.com/

Table 1 An example of the Synonymous Indicators

表 1 同义指标示例

Indicator Name	Synonym	Synonym Sources
B 型钠尿肽	脑尿钠肽	Baidu Encyclopedia
	BNP	Baidu Encyclopedia
	B 型利钠肽	LOINC, Tencent Translation
	B-型利钠肽	Tencent Translation
	B 型钠尿肽	Baidu & iCIBA Translation
	利钠肽 B 型	Baidu & iCIBA Translation
	脑促尿钠排泄肽	iCIBA Translation
	脑利钠肽（物质）	Tencent Translation
	脑钠素	Baidu & iCIBA Translation
	脑钠肽	Baidu & Tencent Translation

2）特征抽取

本文设计了 2 类特征用于指标的二分类，分别是相似度特征和分块打分特征：

①相似度特征

这类特征主要考虑了簇中每一个候选指标与标准指标及其所有同义词的名称相似度和缩写相似度。为了方便描述，以名称相似度为例（缩写相似度也是同理），我们规定簇中候选指标名称为 x^{na}，标准指标名称集合为 $S^{na}=\{s_1^{na},s_2^{na},...,s_n^{na}\}$，其中下标 n 为标准指标及其同义指标的总个数。我们使用以下 4 种相似度来度量：

——最长公共子序列相似度

$$sim_{lcs}(x^{na},S^{na})=\max_i \frac{|lcs(x^{na},s_i^{na})|}{\min(|x^{na}|,|s_i^{na}|)}$$，其中 $|x^{na}|$ 为候选指标名称的字符串长度，$lcs(x^{na},s_i^{na})$ 表示两个指标名称的最大公共子序列。这个相似度可以判定类似上下位关系的指标，比如“血糖”和“血糖（急诊）”在最长公共子序列相似度中为 1。

——Jaccard 相似度

$$sim_{Jac}(x^{na},S^{na})=\max_i \frac{|x^{na}\cap s_i^{na}|}{|x^{na}\cup s_i^{na}|}$$。这个相似度可以判定名称顺序不同的指标，比如“B 型利钠肽”和“利钠肽 B 型”的 Jaccard 相似度为 1。

——余弦相似度

$$sim_{\cos}(x^{na},S^{na})=\max_i \frac{\sum_{k=1}^{d} x_k^{na}, s_{i,k}^{na}}{\sqrt{\sum_{k=1}^{d} x_k^{na^2}}\sqrt{\sum_{k=1}^{d} s_{i,k}^{na^2}}}$$，其中 x^{na} 和 s_i^{na} 均为 multi-hot 形式（0-1 向量中不同的维度表示不同的汉字）。这个相似度衡量的是两个 multi-hot 形式的指标名称的余弦夹角，它受到类似中间插入“-”等格式问题的影响更小一些。

——编辑相似度

$$med(x^{na},S^{na})=\max_i(1-\frac{med(x^{na},s_i^{na})}{\max(|x^{na}|,|s_i^{na}|)})$$，其中 $|x^{na}|$ 是指标名称 x^{na} 的字符串长度，$med(x^{na},s_i^{na})$ 表示由 x^{na} 经插入、替换、删除操作转成 s_i^{na} 所需的最少操作次数。这个相似度衡量的是两个指标名称的编辑距离。

②基于一对多字段的分块打分特征

分块打分特征主要是针对指标参考值这种一对多的字段而言。对于指标参考值来说，由于不同医院对同一个指标，在参考值的上下界设置上有时会略有不同，因此实践中存在着一个指标名称对应多个参考值的现象。为应对这一问题，本文参考文献[36]中的知识库实体对齐分块算法，提出基于参考值的指标分块打分算法。指标分块打分算法基于以下假设：第一，相同的指标拥有相似的参考值；第二，拥有相似参考值的可能就是同一个指标。因此，本文的分块打分算法由两部分组成：首先，为标准指标的每一种参考值寻找一个与之最相似的候选指标参考值；然后，从这些最相似的参考值出发，构建候选指标与标准指标之间的匹配分块。需要注意的是，由于同一个指标可能有多种参考值，算法允许同一个指标出现在不同的块中。本文根据不同块的权重求出候选指标的加权平均得分，以此作为分类特征。

具体来说，给定簇中某一候选指标 x，它所对应的参考值集合为 $X^{ref}=\{x_1^{ref},x_2^{ref},...,x_n^{ref}\}$，其中 x_i^{ref} 表示候选指标 x 的第 i 种参考值范围，以及标准指标（及其同义指标的）参考值集合 $S^{ref}=\{s_1^{ref},s_2^{ref},...,s_m^{ref}\}$，其中 s_i^{ref} 表示标准指标 s 的第 i 种参考值范围。本文定义参考值相似度如下：

定义 3（参考值相似度） 给定两个指标参考值 x^{ref} 和 s^{ref}，定义参考值相似度 $sim_{ref}(x^{ref},s^{ref})=\frac{|x^{ref}\cap s^{ref}|}{|x^{ref}\cup s^{ref}|}$。

对于标准指标的每一个参考值 s_i^{ref}，本文都从簇中找出一个与 s_i^{ref} 最相似的候选指标的参考值 x_j^{ref} 使得 $sim_{ref}(x_j^{ref},s_i^{ref})=\max_k sim_{ref}(x_k^{ref},s_i^{ref})$，并将这两个

指标组成参考值对 p_i^{ref}=(x_j^{ref} , s_i^{ref})。根据参考值对 p_i^{ref}，可以构建指标集对 p_i=(X_i, S_i)，其中 X_i 为所有参考值为 x_j^{ref} 的候选指标的集合，S_i 为所有参考值为 s_i^{ref} 的标准指标及其同义指标的集合。进而定义参考值对相似度如下：

定义 4（参考值对相似度）给定两个参考值对 p_1^{ref} 和 p_2^{ref}，定义参考值对相似度 $sim_{p_ref}(p_1^{ref},p_2^{ref})=\dfrac{sim_{p_cos}(X_1,X_2)+sim_{p_cos}(S_1,S_2)}{2}$，其中 $sim_{p_cos}(X_1,X_2)$ 表示将指标集合 X_1、X_2 表示成 one-hot 形式（0-1 向量中不同的维度表示不同的指标）后两者的余弦相似度。

如图 2 所示，标准参考值 s_1^{ref} 为区间[0, 100]，其最相似的候选参考值 x_1^{ref} 为区间[0, 100]，故其对应指标集对为 p_1=(X_1, S_1)=({A,B},{a,b})。同理，标准参考值 s_2^{ref} =[0,125] 所对应的指标集对 p_2=(X_2, S_2)=({A,B,C},{a,b})。由此，$sim_{p_ref}(p_1^{ref},p_2^{ref})=(\frac{2}{\sqrt{6}}+1)/2=0.9082$。

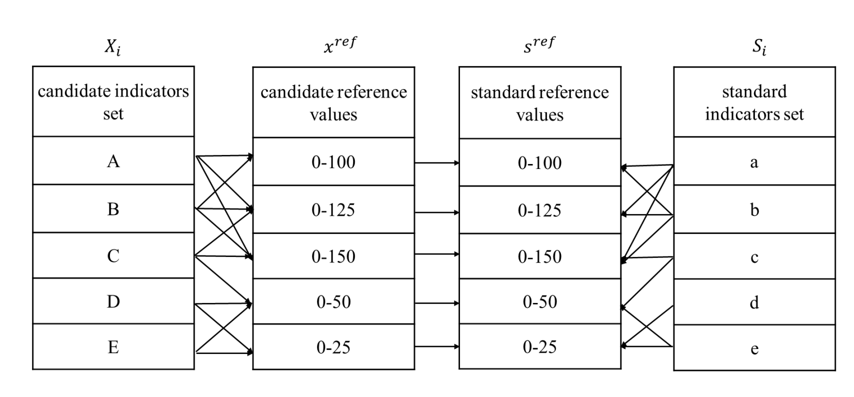


Fig. 2 A schematic diagram about how to calculate similarity of reference value pairs.

图 2 参考值对相似度计算示意图

在匹配分块时，如果两个参考值对的相似度大于阈值θ，即 $sim_{p_ref}(p_1^{ref},p_2^{ref})>\theta$，则其候选指标集合 X_1、X_2 和标准指标集合 S_1、S_2 将被纳入同一个分块中。直观上来说，如果两个参考值对共同拥有的指标越多，它们就越有可能被分为一块。

分块完成后，本文将对每一块做打分处理。定义分块结果 B={B_1, B_2,…, B_n}，其中 n 是块的个数。对于任意一个块 B_i，其得分 $score_i=\alpha+(1-\alpha)\frac{|B_i\cap S'|}{|B_i|}$，其中 $\frac{|B_i\cap S'|}{|B_i|}$ 为块中标准指标所占的比重，S'为所有标准指标的集合，α 是权重参数。块 β_i 中的所有指标共享同一个 $score_i$ 得分。

由于算法允许一个指标出现在不同的块中，因此，一个指标可能会拥有多个不同的分数，需要根据不同块的权重 β_i 求出它的加权平均得分 $score'=\sum_i\beta_i score_i$。作为指标标准化算法的初步尝试，本文简单地认为所有块的权重相同。特别地，如果某指标一个块也没被分入，则其得分为 0。这也是上文计算块中得分 $score_i$ 时进行加权平滑的原因：只要指标能被分入块中，便拥有一个基础得分。

综上，我们就得到了每个指标基于参考值的分块打分特征。

2. 4 重定义簇

簇内二分类将簇内指标划分为标准指标的同义指标与非同义指标。针对非同义指标，本文将其单独取出作为一个新的簇，并从中筛选出一个新的标准名称，继续利用二分类算法进行同义指标的查找，如此迭代进行，直到所有簇内均为同义指标或簇内只剩 1 个指标为止。

2. 5 指标映射与修正

到此阶段，指标标准化算法已进入尾声，只需把簇内的同义指标统一映射为对应的标准指标，并交由医学专业人员对指标的对齐结果进行核验与修正。特别地，聚类过程中可能会把同义的指标分到不同的簇中，二分类过程把簇中非同义指标剔除出来后，人工核验时还需对同义的簇进行合并。

3 实验结果与分析

3.1 数据集

本文从上海临床诊疗信息共享平台中抽取指标数据集进行实验。在指标数据的抽取过程中，本文考虑了两个因素：第一，指标的种类要丰富，否则无法模拟实际应用场景；第二，同义指标的名称要多样化，

否则指标的标准化没有意义。因此，本文以医院为单位，抽取其中所有的指标，保证了丰富性；同时选取了不同指标名称最多的前 8 家医院，以满足多样性。这 8 家医院的不同指标名称数量分别为：1404、1243、1098、1010、992、958、921、849，合并去重后共有 5211 个不同指标名称。在扩充了这些指标名称的缩写字段之后，不同的记录数为 7542 条；在扩充了这些指标名称的缩写和参考值字段之后，不同的记录数达到了 12750 条。在聚类实验部分，本文选择了 236 条数据进行评测。在二分类实验部分，本文以正负例 1:1 的比例进行采样，并将采样结果按 7:3 的比例划分为训练集和测试集，最终得到 947 条训练样本和 406 条测试样本。本文另外选取了 100 个正例和 100 负例作为验证集。

3.2 实验设置

本文通过在验证集上网格搜索，采用参数 $minPts=3$，$\varepsilon=0.35$，阈值 $\theta=0.7$，$\alpha=0.6$ 进行实验，选取梯度上升决策树（gradient boosting decision tree，GBDT）作为最终的二分类模型，并使用 Precision、Recall 和 F1-score 来评价聚类和二分类的效果。

3.3 实验结果

1）聚类算法对比实验

为了考察本文所使用的基于密度的聚类算法（DBSCAN）的有效性，本文选取了四种常见的聚类算法进行对比，它们分别是 k 均值聚类（k-means clustering，K-means）、均值漂移算法（mean shift algorithm，Meanshift）、高斯混合模型（gaussian mixture model，GMM）与凝聚层次聚类（agglomerative hierarchical clustering，AHC）。需要注意的是，由于这四种基准算法除高斯混合模型外都需要事先定义簇数（而本文算法不需要），在实验时本文将它们的聚类数目设为真实的簇数。实验结果如表 2 所示。

Table 2 Comparisons of our method and common clustering methods

表 2 不同聚类算法的性能对比

Clustering Algorithm	Precision	Recall	F1-score
K-means	37.88	21.31	27.27
Meanshift	34.93	18.85	24.49
GMM	**42.17**	23.98	30.58
AHC	35.16	20.30	25.74
Our DBSCAN	27.85	**91.36**	**42.68**

从表中可以看出，本文基于密度的聚类算法的 F1-score 明显高于其它 4 种聚类算法，其提高幅度均在 10%以上。然而，虽然本文方法的 Recall 能达到 91.36%，但 Precision 仍然不是很高，这也显示了本文在聚类后进一步进行二分类映射的必要性。

2）二分类算法对比实验

①不同分类特征和不同分类器的对比

为了考察不同分类特征和不同分类器对分类性能的影响，本文选择不同的特征组合，将它们在逻辑回归（logistic regression，LR）、朴素贝叶斯（naive bayes，NB）、 k 近邻（k-nearest neighbor，KNN）、支持向量机（support vector machine，SVM）、随机森林（random forest，RF）、梯度上升决策树（gradient boosting decison tree，GBDT）等不同分类器下的 F1-score 进行对比。实验结果如表 3 所示，其中特征字段的名称（name）、缩写（abbreviation，Abbr.）和参考值（reference value，Ref.）分别表示名称相似度特征、缩写相似度特征和参考值分块打分特征。

Table 3 Comparisons of different classification features and different classifiers

表 3 不同分类算法的性能对比

Features	LR	NB	KNN	SVM	RF	GBDT
Name	76.56	74.59	76.58	75.26	76.17	76.96
Abbr.	74.24	73.63	73.95	74.16	77.64	77.25
Ref.	74.09	70.38	75.83	53.96	77.92	78.71
Name+Abbr.	79.10	77.67	78.82	78.14	83.03	81.05
Name+Ref.	78.55	75.86	76.50	75.90	82.60	82.45
Abbr.+Ref.	77.11	74.94	76.03	74.44	80.31	80.83
Name+Abbr.+Ref.	79.30	78.55	78.05	78.47	83.94	**85.27**

从表中可以看出，当使用名称相似度特征、缩写相似度特征和参考值分块打分特征，辅以 GBDT 分类器时，分类效果最好，其 F1 值可达 85.27%。从表中横向来看，无论使用哪种特征，大部分情况下都是 GBDT 分类效果最好，而 NB 分类效果最差。这是因为 GBDT 使用 Boosting 方法进行集成学习，能够有效提高泛化性能，而 NB 分类器的条件独立假设在本文中很难成立。从表中纵向来看，无论哪种分类器，

基本都是随着特征数目的增多，分类效果越来越好，当使用全部三类分类特征时，分类效果达到最好。

②与现有方法的对比

最后，本文还从最近三年来发表的实体对齐方法中选择了 3 种 state-of-the-art 方法，与本文所使用全部三类特征辅以 GBDT 分类器的二分类方法进行对比，这 3 种基准方法分别是：

知识图谱融合方法（KG Fusion）：Wang 等人[37]设计不同类型的属性相似度，使用机器学习方法进行多源知识图谱的融合。

诊断对齐方法（Diag. Alignment）：Ning 等人[38]利用诊断的上下位信息和属性相似度将中文诊断映射为 ICD 编码。

知识库对齐方法（KB Alignment）：王雪鹏等人[39]利用网络语义标签进行多元知识库的实体对齐。

需要注意的是，由于本文任务中既没有属性信息，又没有上下文信息，所以在实际实验中 3 种基准方法的部分特征没法使用，而主要使用了其中的实体名称和缩写的相似度计算方法。

与现有方法的对比实验结果如表 4 所示。从表中可以看出，本文方法在所有方法中取得了最好的分类结果，其 Precision、Recall 和 F1-score 分别为 86.84%、83.76%和 85.27%。值得注意的是，对比表 3 最后一列，当使用 GBDT 分类器时，本文方法的任意两类特征组合的 F1-score 都比现有方法来得好。这是因为本文的算法专门针对检验检测指标进行设计，因而能取得更好的效果。

Table 4 Performance comparison of entity alignment

表 4 与现有方法的对比

Method	Precision	Recall	F1-score
KG Fusion	79.23	73.60	76.32
Diag. Alignment	81.67	74.62	77.98
KB Alignment	**87.20**	72.59	79.22
Ours	86.84	**83.76**	**85.27**

4 结论

本文针对区域医疗健康平台中的检验检查指标标准化，先根据指标的字面特征进行聚类，再使用相似度特征和分块打分特征迭代地进行二分类映射。实验表明，最终的二分类映射，其 F1-score 可以达到 85.27%，优于现有方法。在未来，可以将指标的同义词信息及参考值信息应用到聚类算法之中，并尝试使用更多的相似性度量特征，以获得更好的结果。

5 致谢

感谢上海中医药大学的张海涛同学和同济大学医学院的李阳同学在数据集标注上提供的帮助。

参 考 文 献

[1] Mining W I D. Data Mining: Concepts and Techniques[J]. Morgan Kaufinann, 2006.

[2] Vapnik V. The nature of statistical learning theory[M]. Springer science & business media, 2013.

[3] Han X, Sun L. A generative entity-mention model for linking entities with knowledge base[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 945-954.

[4] Zhang W, Sim Y C, Su J, et al. Entity linking with effective acronym expansion, instance selection, and topic modeling[C]//IJCAI. 2011, 2011: 1909-1914.

[5] Newcombe H B, Kennedy J M, Axford S J, et al. Automatic linkage of vital records[J]. Science, 1959: 954-959.

[6] Fellegi I P, Sunter A B. A theory for record linkage[J]. Journal of the American Statistical Association, 1969, 64(328): 1183-1210.

[7] Cochinwala M, Kurien V, Lalk G, et al. Efficient data reconciliation[J]. Information Sciences, 2001, 137(1-4): 1-15.

[8] Elfeky M G, Verykios V S, Elmagarmid A K. TAILOR: A record linkage toolbox[C]//Data Engineering, 2002.

[9] Christen P. Automatic training example selection for scalable unsupervised record linkage[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, 2008: 511-518.

[10] Kantardzic M. Data mining: concepts, models, methods, and algorithms[M]. John Wiley & Sons, 2011.

[11] Chen Z, Kalashnikov D V, Mehrotra S. Exploiting context analysis for combining multiple entity resolution systems[C]//Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009: 207-218.

[12] Cohen W W, Richman J. Learning to match and cluster large high-dimensional data sets for data integration[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: 475-480.

[13] McCallum A, Wellner B. Conditional models of identity uncertainty with application to noun coreference[C]//Advances in neural information processing systems. 2005: 905-912.

[14] Pasula H, Marthi B, Milch B, et al. Identity uncertainty and citation matching[C]//Advances in neural information processing systems. 2003: 1425-1432.

[15] Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: 269-278.

[16] Tejada S, Knoblock C A, Minton S. Learning domain-independent string transformation weights for high accuracy object identification[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: 350-359.

[17] Arasu A, Götz M, Kaushik R. On active learning of record matching packages[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of data. ACM, 2010: 783-794.

[18] Bhattacharya I, Getoor L. A latent dirichlet model for unsupervised entity resolution[C]//Proceedings of the 2006 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2006: 47-58.

[19] Hall R, Sutton C, McCallum A. Unsupervised deduplication using cross-field dependencies[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008: 310-317.

[20] Domingos P. Multi-relational record linkage[C]//In Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining. 2004.

[21] Singla P, Domingos P. Entity resolution with markov logic[C]//Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE, 2006: 572-582.

[22] Rastogi V, Dalvi N, Garofalakis M. Large-scale collective entity matching[J]. Proceedings of the VLDB Endowment, 2011, 4(4): 208-218.

[23] Blanco R, Ottaviano G, Meij E. Fast and space-efficient entity linking for queries[C]//Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015: 179-188.

[24] Shen W, Wang J, Luo P, et al. Linking named entities in tweets with knowledge base via user interest modeling[C]//Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 68-76.

[25] Han X, Sun L, Zhao J. Collective entity linking in web text: a graph-based method[C]//Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011: 765-774.

[26] Gentile A L, Zhang Z, Xia L, et al. Graph-based semantic relatedness for named entity disambiguation[J]. 2009.

[27] Alhelbawy A, Gaizauskas R. Graph ranking for collective named entity disambiguation[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2014, 2: 75-80.

[28] Hoffart J, Yosef M A, Bordino I, et al. Robust disambiguation of named entities in text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 782-792.

[29] He Z, Liu S, Li M, et al. Learning entity representation for entity disambiguation[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2013, 2: 30-34.

[30] Huang H, Heck L, Ji H. Leveraging deep neural networks and knowledge graphs for entity disambiguation[J]. arXiv preprint arXiv:1504.07678, 2015.

[31] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[32] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 160-167.

[33] Ester M, Kriegel H P, Xu X. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise[C]// International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996:226-231.

[34] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth.[J]. Studies in Health Technology & Informatics, 2006, 121(121):279.

[35] Mcdonald C J, Huff S M, Suico J G, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update.[J]. Clinical Chemistry, 2003, 49(4):624.

[36] Zhuang Y, Li G, Zhong Z, et al. Hike: A Hybrid Human-Machine Method for Entity Alignment in Large-Scale Knowledge Bases[C]//Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, 2017: 1917-1926.

[37] Wang H, Fang Z, Zhang L, et al. Effective online knowledge graph fusion[C]//International Semantic Web Conference. Springer, Cham, 2015: 286-302.

[38] Ning W, Yu M, Zhang R. A hierarchical method to automatically encode Chinese diagnoses through semantic similarity estimation[J]. BMC medical informatics and decision making, 2016, 16(1): 30.

[39] X.-P. Wang, K. Liu, S.-Z. He, S.-L. Liu, Y.-Z. Zhang, and J. Zhao, “Multi-source knowledge bases entity alignment by leveraging semantic tags,” Jisuanji Xuebao/Chinese Journal of Computers, vol. 40, no. 3, pp. 701 – 711, 2017.(in Chinese)
（王雪鹏, 刘康, 何世柱, 等. 基于网络语义标签的多源知识库实体对齐算法[J]. 计算机学报, 2017, 40(3): 701r711.）